Genomic and transcriptional aberrations linked to breast cancer pathophysiologies

Koei Chin,[1,5] Sandy DeVries,[1,5] Jane Fridlyand,[1,5] Paul T. Spellman,[2] Ritu Roydasgupta,[1] Wen-Lin Kuo,[1,2] Anna Lapuk,[1,2] Richard M. Neve,[1,2] Zuwei Qian,[4] Tom Ryder,[4] Fanqing Chen,[2] Heidi Feiler,[1,2] Taku Tokuyasu,[1] Chris Kingsley,[1] Shanaz Dairkee,[3] Zhenhang Meng,[3] Karen Chew,[1] Daniel Pinkel,[1] Ajay Jain,[1] Britt Marie Ljung,[1] Laura Esserman,[1] Donna G. Albertson,[1] Frederic M. Waldman,[1,6] and Joe W. Gray[1,2,6,*]

[1] Comprehensive Cancer Center, 2340 Sutter Street, University of California, San Francisco, San Francisco, California 94143
[2] Life Sciences Division, Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, California 94127
[3] California Pacific Medical Center, 475 Brannan Street, San Francisco, California 94107
[4] Affymetrix, Inc., 3450 Central Expressway, Santa Clara, California 95051
[5] These authors contributed equally to this work.
[6] These authors contributed equally to this work.
*Correspondence: jwgray@lbl.gov

Summary

This study explores the roles of genome copy number abnormalities (CNAs) in breast cancer pathophysiology by identifying associations between recurrent CNAs, gene expression, and clinical outcome in a set of aggressively treated early-stage breast tumors. It shows that the recurrent CNAs differ between tumor subtypes defined by expression pattern and that stratification of patients according to outcome can be improved by measuring both expression and copy number, especially high-level amplification. Sixty-six genes deregulated by the high-level amplifications are potential therapeutic targets. Nine of these (*FGFR1*, *IKBKB*, *ERBB2*, *PROCC*, *ADAM9*, *FNTA*, *ACACA*, *PNMT*, and *NR1D1*) are considered druggable. Low-level CNAs appear to contribute to cancer progression by altering RNA and cellular metabolism.

Introduction

It is now well established that breast cancers progress through accumulation of genomic (Albertson et al., 2003; Knuutila et al., 2000) and epigenomic (Baylin and Herman, 2000; Jones, 2005) aberrations that enable the development of aspects of cancer pathophysiology such as reduced apoptosis, unchecked proliferation, increased motility, and increased angiogenesis (Hanahan and Weinberg, 2000). Discovery of the genes that contribute to these pathophysiologies when deregulated by recurrent aberrations is important to understanding mechanisms of cancer formation and progression and to guide improvements in cancer diagnosis and treatment.

Analyses of expression profiles have been particularly powerful in identifying distinctive breast cancer subsets that differ in biological characteristics and clinical outcome (Perou et al., 1999, 2000; Sorlie et al., 2001, 2003). For example, unsupervised hierarchical clustering of microarray-derived expression data has identified intrinsically variable gene sets that distinguish five breast cancer subtypes—basal-like, luminal A, luminal B, ERBB2, and normal breast-like. The basal-like and ERBB2 subtypes have been associated with strongly reduced survival durations in patients treated with surgery plus radiation (Perou et al., 2000; Sorlie et al., 2001), and some studies have suggested that reduced survival duration in poorly performing subtypes is caused by an inherently high propensity to metastasize (Ramaswamy et al., 2003). These analyses already have led to the development of multigene assays that stratify patients into groups that can be offered treatment strategies based on risk of progression (Esteva et al., 2005; Gianni et al., 2005; van 't Veer et al., 2002; van de Vijver et al., 2002). However, the predictive power of these assays is still not as high as desired, and the assays have not been fully tested in patient populations treated with aggressive adjuvant chemotherapies.

Analyses of breast tumors using fluorescence in situ hybridization (Al-Kuraya et al., 2004; Kallioniemi et al., 1992; Press

SIGNIFICANCE

This study indicates that the accuracy with which breast patients can be stratified according to outcome can be improved by combining analyses of gene expression and genome copy number. Markers for high-level amplification and/or overexpression of genes at 8p11, 11q13, 17q12, and/or 20q13 are particularly strong predictors of reduced survival duration. Genes in these regions are high-priority therapeutic targets for treatment of patients that respond poorly to current aggressive therapies. The statistically significant deregulation of genes involved in RNA and cellular metabolism by low-level CNAs suggests that these events contribute to breast cancer progression by increasing basal metabolism.

et al., 2005; Tanner et al., 1994) and comparative genomic hybridization (Kallioniemi et al., 1994; Loo et al., 2004; Naylor et al., 2005; Pollack et al., 1999) show that breast tumors also display a number of recurrent genome copy number aberrations, including regions of high-level amplification that have been associated with adverse outcome (Al-Kuraya et al., 2004; Cheng et al., 2004; Isola et al., 1995; Jain et al., 2001; Press et al., 2005). This raises the possibility of improved patient stratification through combined analysis of gene expression and genome copy number (Barlund et al., 2000; Pollack et al., 2002; Ray et al., 2004; Yi et al., 2005). In addition, several studies of specific chromosomal regions of recurrent abnormality at 17q12 (Kauraniemi et al., 2001, 2003) and 8p11 (Gelsi-Boyer et al., 2005; Ray et al., 2004) show the value of combined analysis of genome copy number and gene expression for identification of genes that contribute to breast cancer pathophysiology by deregulating gene expression.

We have extended these studies by performing combined analyses of genome copy number and gene expression to identify genes that contribute to breast cancer pathophysiology, with emphasis on those that are associated with poor response to current therapies. By associating clinical endpoints with genome copy number and gene expression, we showed strong associations between expression subtype and genome aberration composition, and we identified four regions of recurrent amplification associated with poor outcome in treated patients. Gene expression profiling revealed 66 genes in these regions of amplification whose expression levels were deregulated by the high-level amplifications. We also found a surprising association between low-level CNAs and upregulation of genes associated with RNA and protein metabolism that may suggest a mechanism by which these aberrations contribute to cancer progression.

Results

We assessed genome copy number using BAC array CGH (Hodgson et al., 2001; Pinkel et al., 1998; Snijders et al., 2001; Solinas-Toldo et al., 1997) and gene expression profiles using Affymetrix U133A arrays (Ramaswamy et al., 2003; Reyal et al., 2005) in breast tumors from a cohort of patients treated according to the standard of care between 1989 and 1997 (surgery, radiation, hormonal therapy, and treatment with high-dose adriamycin and cytoxan as indicated). We measured genome copy number profiles for 145 primary breast tumors and gene expression profiles for 130 primary tumors, of which 101 were in common. We analyzed these data to identify recurrent genomic and transcriptional abnormalities, and we assessed associations with clinical endpoints to identify genomic events that might contribute to cancer pathophysiology.

Molecular characteristics and associations

Genome copy number and gene expression features

We found that the recurrent genome copy number and gene expression characteristics measured for the patient cohort in this study were similar to those reported in earlier studies. We summarize these briefly.

Figures 1A and 1B show numerous regions of recurrent genome CNA and nine regions of recurrent high-level amplification involving regions of chromosomes 8, 11, 12, 17, and 20, while Figure 2 shows that analysis of these data using unsupervised hierarchical clustering resolves these tumors into the ''1q/16q'' (or ''simple''), ''complex,'' and ''amplifier'' genome aberration subtypes (Fridlyand et al., 2006). The genomic extents of the regions of amplification are listed in Table 1. These were generally similar to those reported in earlier studies using chromosome (Kallioniemi et al., 1994) and array CGH (Loo et al., 2004; Naylor et al., 2005; Pollack et al., 1999, 2002). Several of these regions of amplification were frequently coamplified. Declaring a Fisher exact test p value of less than 0.05 for pairwise associations to be suggestive of possible significant coamplification, we found coamplification of 8q24 and 20q13 and coamplification of regions at 11q13-14, 12q13-14, 17q11-12, and 17q21-24. These analyses were underpowered to achieve significance with proper correction for multiple testing, so these associations are suggestive but not significant. However, these associations were consistent with the report of Al-Kuraya et al. (2004), who showed evidence for coamplification of genes in several of these regions of amplification including *ERBB2*, *MYC*, *CCND1*, and *MDM2*, and that of Naylor et al. (2005) showing coamplification of 17q12 and 17q25.

Figure S1 (in the Supplemental Data available with this article online) shows that unsupervised hierarchical clustering of intrinsically variable genes resolves the tumors in our study cohort into the luminal A, luminal B, basal-like, and ERBB2 expression subtypes previously reported for breast tumors (Perou et al., 1999, 2000; Sorlie et al., 2003). We assessed the genomic characteristics of these expression subtypes in subsequent analyses.

Associations between CNAs and expression

Combined analyses of genome copy number and expression showed that the recurrent genome CNAs differed between expression subtypes and identified genes whose expression levels were significantly deregulated by the CNAs. Figures 1C–1J show the recurrent CNAs for each expression subtype. In these analyses, we assigned each tumor to the expression subtype cluster (basal-like, ERBB2, luminal A, and luminal B) to which its expression profile was most highly correlated. We did not assess aberrations in normal-like tumors due to the small number of such tumors. Figure 1C shows that the basal-like tumors were relatively enriched for low-level copy number gains involving 3q, 8q, and 10p and losses involving 3p, 4p, 4q, 5q, 12q, 13q, 14q, and 15q, while Figure 1D shows that high-level amplification at any locus was infrequent in these tumors. Figure 1E shows that ERBB2 tumors were relatively enriched for increased copy number at 1q, 7p, 8q, 16p, and 20q and reduced copy number at 1p, 8p, 13q, and 18q. Figure 1F shows that amplification of *ERBB2* was highest in the ERBB2 subtype as expected, but amplification of noncontiguous, distal regions of 17q also was frequent as previously reported (Barlund et al., 1997). Figure 1G shows that increased copy number at 1q and 16p and reduced copy number at 16q were the most frequent abnormalities in luminal A tumors, while Figure 1H shows that high-level amplifications at 8p11-12, 11q13-14, 12q13-14, 17q11-12, 17q21-24, and 20q13 were relatively common in this subtype. Figure 1I shows that gains of chromosomes 1q, 8q, 17q, and 20q and losses involving portions of 1p, 8p, 13q, 16q, 17p, and 22q were prevalent in luminal B tumors, while Figure 1J shows that high-level amplifications involving 8p11-12, two regions of 8q, and 11q13-14 were frequent. Bergamaschi et al. (2006) have reported similar CNA patterns for the luminal A, luminal B, basal, and ERBB2 expression clusters.

In order to understand how the genome aberrations influence cancer pathophysiologies, we identified genes that were

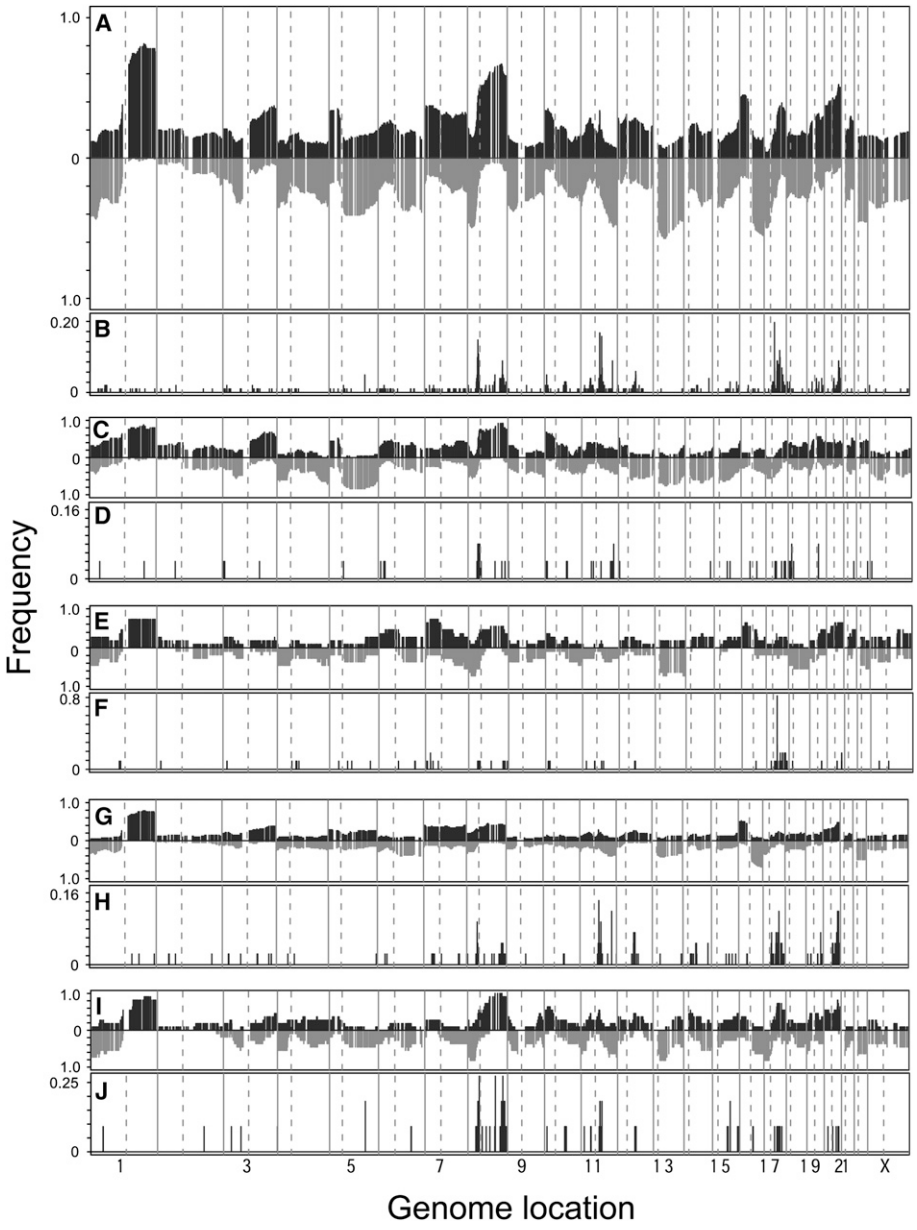

Figure 1. Recurrent abnormalities in 145 primary breast tumors

A: Frequencies of genome copy number gain and loss plotted as a function of genome location with chromosomes 1pter to the left and chromosomes 22qter and X to the right. Vertical lines indicate chromosome boundaries, and vertical dashed lines indicate centromere locations. Positive and negative values indicate frequencies of tumors showing copy number increases and decreases, respectively, with gain and loss as described in the Experimental Procedures.

B: Frequencies of tumors showing high-level amplification. Data are displayed as described in **A**.

C–J: Frequencies of tumors showing significant copy number gains and losses as defined in **A** (upper member of each pair) or high-level amplifications as defined in **B** (lower member of each pair) in tumor subtypes defined according to expression phenotype; **C** and **D**, basal-like; **E** and **F**, ERBB2; **G** and **H**, luminal A; **I** and **J**, luminal B. Data are displayed as described in **A**.

deregulated by recurrent genome CNAs. We took these genes to be those whose expression levels were significantly associated with copy number (Holm-adjusted p value < 0.05). These genes, which represent about 10% of the genome interrogated by the Affymetrix HGU133A arrays used in this study, and their copy number-expression level correlation coefficients are listed in Table S3. This extent of genome-aberration-driven deregulation of gene expression is similar to that reported in earlier studies (Hyman et al., 2002; Pollack et al., 1999). We tested associations between copy number and expression level for 186 genes in regions of amplification at 8p11-12, 11q13-q14, 17q11-12, and 20q13, and we identified 66 genes in these regions whose expression levels were correlated with copy number (FDR < 0.01, Wilcoxon rank-sum test; Table 3). These genes define the transcriptionally important extents of the regions of recurrent amplification. Twenty-three were from a 5.5 Mbp region at 8p11-12 flanked by *SPFH2* and *LOC441347*, ten were from a 6.6 Mbp region at 11q13-14 flanked by *CCND1* and *PRKRIR*, nineteen were from a 3.1 Mbp region at 17q12 flanked by *LHX1* and *NR1D1*, and fourteen were from a 5.4 Mbp region at 20q13 flanked by *ZNF217* and *C20orf45*.

Since the recurrent genome aberrations differed between expression subtypes, we explored the extent to which the expression subtypes were determined by genome copy number. Specifically, we applied unsupervised hierarchical clustering to intrinsically variable genes *after removing genes whose expression levels were correlated with copy number*. Figure 4 shows that the tumors still resolve into the basal-like and luminal classes. However, the ERBB2 cluster was lost.

Associations with clinical variables

Associations with histopathology

Figure 2 and Table 2 summarize associations of histopathological features with aspects of genome abnormality, including

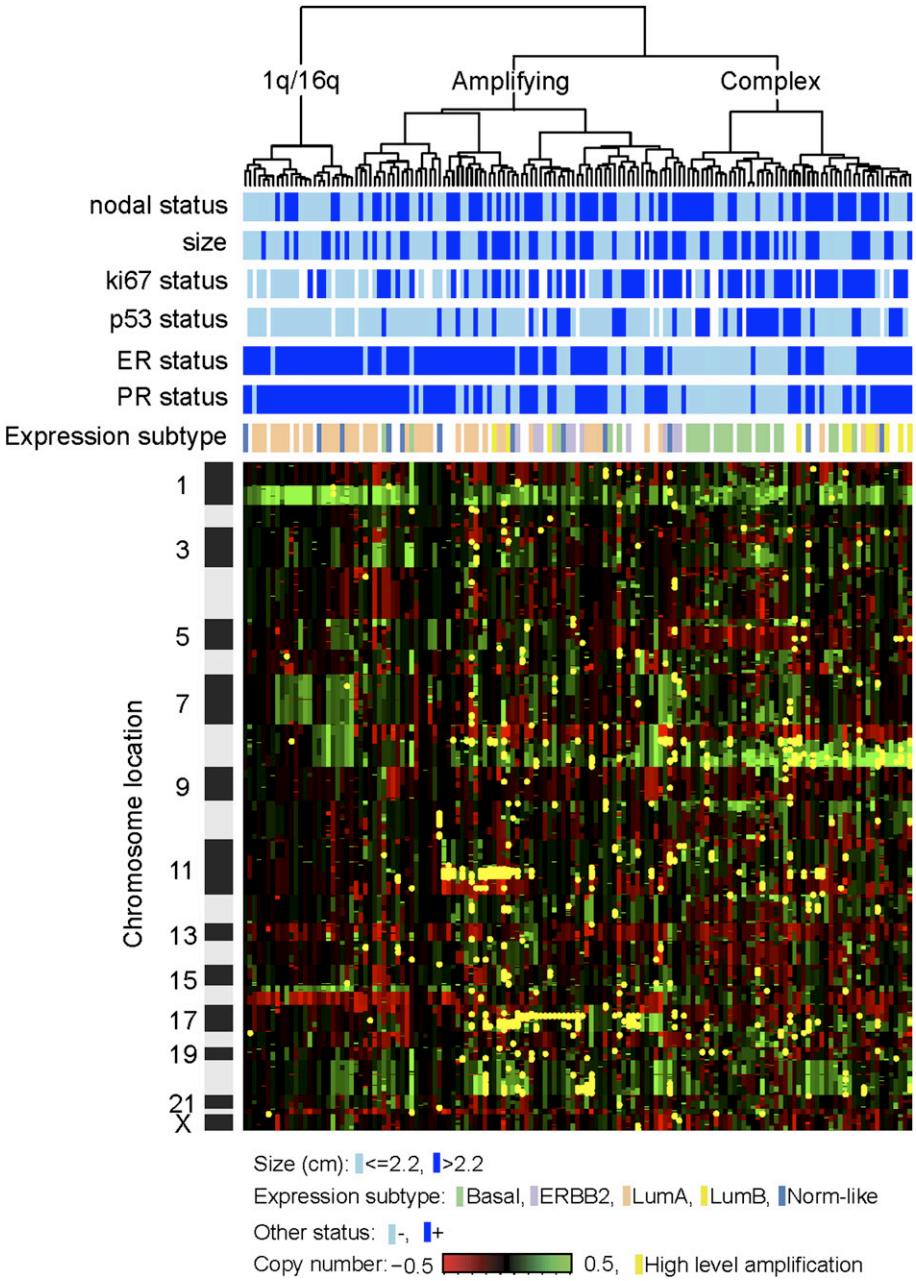

Figure 2. Unsupervised hierarchical clustering of genome copy number profiles measured for 145 primary breast tumors

Green indicates increased genome copy number, and red indicates decreased genome copy number. The three major genomic clusters from left to right are designated 1q/16q, complex, and amplifying. The bar to the left indicates chromosome locations with chromosome 1pter to the top and 22qter and X to the bottom. The locations of the odd-numbered chromosomes are indicated. The upper color bars indicate biological and clinical aspects of the tumors. Color codes are indicated at the bottom of the figure. Dark blue indicates positive status, and light blue indicates negative status for node, ER, PR, and p53 expression. For Ki67, dark blue indicates fraction >0.1, and light blue indicates fraction <0.1. For size, light blue indicates size <2.2 cm, and dark blue indicates size >2.2 cm. Color codes for the expression bar are as follows: orange, luminal A; dark blue, normal breast-like; light blue, ERBB2; green, basal-like; yellow, luminal B.

recurrent genome abnormalities, total number of copy number transitions, fraction of the genome altered (FGA), number of chromosomal arms containing at least one amplification, number of recurrent amplicons, and presence of at least one recurrent amplification. These analyses showed that ER/PR-negative tumors were predominantly found in the basal-like expression and "complex" genome aberration subtypes, respectively. Node-positive tumors had significantly more amplified arms and recurrent amplicons than node-negative samples but showed a much more moderate difference in terms of low-level copy number transitions. Stage 1 tumors had moderately fewer low- and high-level changes than higher-stage tumors. The number of low- and high-level abnormalities increased with SBR grade. Interestingly, the "complex" tumors showing many low-level abnormalities were more strongly associated with aberrant p53 expression than "amplifying" tumors. "Simple" tumors tended to have Ki67 proliferation indices <10%, while "complex" and "amplifying" tumors typically had Ki67 indices >10%. The number of amplifications increased significantly with tumor size, but the number of low-level changes did not. We observed no association of genomic changes with the age at diagnosis.

Associations with outcome

Figure 2 and Table S2 summarize associations between histopathological, transcriptional, and genomic characteristics and outcome endpoints identified using multivariate regression analysis. Histopathological features including size and nodal status were significantly associated with survival duration and/or disease recurrence in univariate analyses (Table S1) and were included in the multivariate regressions described below.

The tumor subtypes based on patterns of gene expression or genome aberration content showed moderate associations with

Table 1. Univariate and multivariate associations for individual amplicons and/or disease-specific survival and distant recurrence

Amplicon	Flanking clone (left)	Flanking clone (right)	Kb start	Kb end	p value, univariate		p value, luminal A, univariate		p value, multivariate	
					survival	recurrence	survival	recurrence	survival	recurrence
8p11-12	RP11-258M15	RP11-73M19	33579	43001	0.011	0.004	0.022	0.004	0.037	0.006
8q24	RP11-65D17	RP11-94M13	127186	132829	0.830	0.880	0.140	1.0	0.870	0.720
11q13-14	CTD-2080I19	RP11-256P19	68482	71659	0.540	0.410	0.016	0.240	0.660	0.440
11q13-14	RP11-102M18	RP11-215H8	73337	78686	0.230	0.150	0.016	0.240	0.360	0.190
12q13-14	BAL12B2624	RP11-92P22	67191	74053	0.250	0.260	0.230	0.098	0.920	0.960
17q11-12	RP11-58O8	RP11-87N6	34027	38681	0.004	0.004	1.0	1.0	0.022	0.008
17q21-24	RP11-234J24	RP11-84E24	45775	70598	0.960	0.920	0.610	0.290	0.530	0.630
20q13	RMC20B4135	RP11-278I13	51669	53455	0.340	0.800	0.048	0.140	0.590	0.970
20q13	GS-32I19	RP11-94A18	55630	59444	0.087	0.230	0.048	0.140	0.060	0.220
Any amplicon					0.005	0.003	0.024	0.120	0.034	0.009

Also shown are the chromosomal positions of the beginning and ends of the amplicons and the flanking clones. Associations are shown for the entire sample set and for luminal A tumors (univariate associations only).

outcome endpoints. For example, Figure 3A shows that patients with tumors classified as ERBB2 based on expression pattern had significantly shorter disease-specific survival than patients classified as luminal A or luminal B as previously reported (Perou et al., 2000; Sorlie et al., 2001). Unlike these earlier reports, patients with tumors classified as basal-like did not do significantly worse than patients with luminal or normal breast-like tumors, although there was a trend in that direction. In addition, Figure 3B indicates that patients with tumors classified as "1q/16q" based on genome aberration content tended to have longer disease-specific survival than patients with "complex" or "amplifier" tumors.

We found that high-level amplification was most strongly associated with poor outcome in this aggressively treated patient population. Amplification at any of the nine recurrent amplicons was an independent risk factor for reduced survival duration ($p < 0.04$) and distant recurrence ($p < 0.01$) in a multivariate Cox-proportional model that included tumor size and nodal status. Figure 3C, for example, shows that patients whose tumors had at least one recurrent amplicon survived a significantly shorter time than did patients with tumors showing no amplifications. More specifically, amplifications of 8p11-12 or 17q11-12 (*ERBB2*) were significantly associated with disease-specific survival and distant recurrence in all patients in multivariate regressions (Table 1). Importantly, we found that stratification according to amplification status allowed identification of patients with poor outcome even within an expression subtype. Figure 3D, for example, shows that patients with luminal A tumors and amplification at 8p11-12, 11q13-14, or 20q13 had significantly shorter disease-specific survival than patients without amplification in one of these regions (the number of samples in the luminal A subtype group was too small for multivariate regressions). Amplification at 8p11-12 was most strongly associated with distant recurrence in the luminal A subtype.

Considering the strong association between amplification and outcome, we explored the possibility that some of these genes were overexpressed in tumors in which they were not amplified and that overexpression was associated with reduced survival duration in those tumors. Increased expression levels of seven genes (see Table 3) were associated with reduced survival or distant recurrence at the $p < 0.1$ level, but only two, the growth factor receptor-binding protein GRB7 (17q) and the keratin-associated protein KTRAP5-9 (11q), at the $p < 0.05$ level. Interestingly, this analysis also revealed an unexpected association between *reduced* expression levels of genes from regions of amplification and poor outcome (either disease-free survival or distant recurrence) in tumors without relevant amplifications ($p < 0.05$). This was especially prominent for genes from the region of amplification at 8p11-12 (14 of 23 genes in this region showed this association), while only two genes from regions of adverse-outcome-associated amplifications on chromosomes 17q and 20q showed this association. Following this lead, we tested associations between outcome and reduced copy number at 8p11-12 in patients in tumors in which 8p11-12 was not

Table 2. Associations of genomic variables with clinical features

	Fraction of genome altered[1]	Total number of transitions[1]	Number of amplified arms[1]	Number of recurrent amplicons[1]	Presence of recurrent amplicons[2]
1. ER (negative versus positive)	<0.001	<0.001	0.376	0.147	0.482
2. PR (negative versus positive)	0.005	<0.001	<0.050	0.319	0.390
3. Nodes (positive versus negative)	0.053	0.106	0.012	0.012	0.008
4. Stage (>1 versus 1)	0.013	0.052	0.045	0.312	0.368
5. ERBB2 (positive versus negative)	0.650	0.830	0.015	<0.001	<0.001
6. Ki67 (>0.1 versus <0.1)	0.013	0.031	0.024	0.010	0.005
7. P53 (positive versus negative)	0.001	<0.001	0.043	0.573	0.171
8. Size	0.339	0.088	0.016	0.005	0.015
9. Age at Dx	0.767	0.361	0.223	0.905	0.947
10. SBR grade	<0.001	<0.001	0.008	0.206	0.035
11. Expression subtype	<0.001	<0.001	0.002	0.003	<0.001
12. Genomic subtype	<0.001	<0.001	<0.001	<0.001	<0.001

[1]Kruskal-Wallis test (1–7, 11, and 12), significance of robust linear regression standardized coefficient (8–10).
[2]Fisher exact test (1–7, 11, and 12), significance of robust linear regression standardized coefficient (8–10).

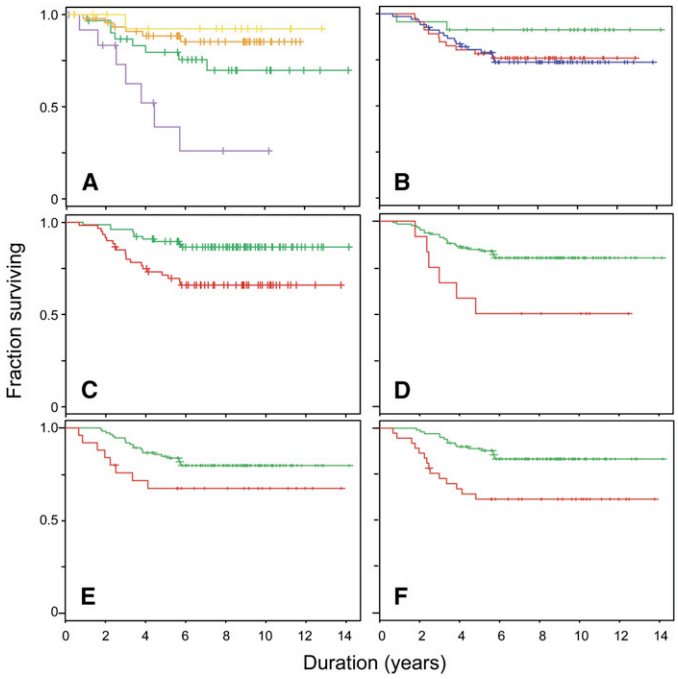

Figure 3. Kaplan-Meyer plots showing survival in breast tumor subclasses

A: Disease-specific survival in 130 breast cancer patients whose tumors were defined using expression profiling to be basal-like (green curve), luminal A (yellow curve), luminal B (orange curve), and ERBB2 (purple curve) class.

B: Disease-specific survival of patients with tumors classified by genome copy number aberration analysis as 1q/16q (green), complex (red), and amplifying (blue).

C: Survival of patients with (red curve) and without (green curve) amplification at any region of recurrent amplification.

D: Survival of patients whose tumors were defined using expression profiling to be luminal A tumors with (red curve) and without (green curve) amplification at 8p11-12, 11q13, and/or 20q.

E: Survival of patients whose tumors were not amplified at 8p11-12 and had normal (green curve) or reduced (red curve) genome copy number at 8p11-12.

F: Survival of patients whose tumors had normal (green curve) or abnormal (red curve) genome copy number at 8p11-12.

amplified. Figure 3E shows that patients with reduced copy number at 8p11-12 did worse than patients without a deletion in this region. Figure 3F shows that patients in the overall study with high-level amplification or deletion at 8p11-12 survived significantly shorter survival ($p = 0.0017$) than patients without either of those events.

We also tested for associations of low-level genome copy number changes with the outcome endpoints. The most frequent low-level copy number changes (e.g., increased copy number at 1q, 8q, and 20q or decreased copy number at 16q) were not significantly associated with outcome endpoints. However, we did find a significant association of the loss of a small region on 9q22 with adverse outcome, both disease-specific survival and distal recurrence, which persisted even after correction for multiple testing ($p < 0.05$, multivariate Cox regression). This region is defined by BACs, CTB-172A10, and RP11-80F13. We also found a marginally significant association between fraction of the genome lost and disease-specific survival in luminal A tumors ($p < 0.02$ and < 0.06 for univariate and multivariate regression, respectively, Cox-proportional regression).

We used the program GoStat (Beissbarth and Speed, 2004) to identify the Gene Ontology (GO) classes of 1444 unique genes (1734 probe sets) whose expression levels were preferentially modulated by low-level CNAs compared to 3026 probe sets whose expression levels did not show associations with copy number. The GO categories most significantly overrepresented in the set of genes with a dosage effect compared to genes with no or minimal dosage effect involved RNA processing (Holm adjusted p value < 0.001), RNA metabolism ($p < 0.01$), and cellular metabolism ($p < 0.02$).

Discussion

This paper describes a comprehensive analysis of gene expression and genome copy number in aggressively treated primary human breast cancers performed in order to (1) identify genomic events that can be assayed to better stratify patients according to clinical behavior, (2) develop insights into how molecular aberrations contribute to breast cancer pathogenesis, and (3) discover genes that might be therapeutic targets in patients that do not respond well to current therapies. An accompanying paper in this issue of *Cancer Cell* shows that many of these aberrations are found in subsets of breast cancer cell lines that can be manipulated to confirm functions suggested by associations with pathophysiology established here (Neve et al., 2006).

Molecular markers that predict outcome

Our combined analyses of genome copy number and gene expression focused on tumors from patients treated more aggressively than those in previously published studies (Perou et al., 2000; Sorlie et al., 2001) (i.e., with surgery, radiation of the surgical margins, hormonal therapy for ER-positive disease, and aggressive adjuvant chemotherapy as indicated) and revealed two important associations.

First, they showed that the survival of patients with tumors classified as basal-like according to expression pattern did not have significantly worse outcome than patients with luminal or normal-like tumors in this tumor set, unlike previous reports (Perou et al., 2000; Sorlie et al., 2001) (see Figure 3A), although there was a trend toward lower survival. However, patients with ERBB2-positive tumors did have significantly increased death from disease and shorter recurrence-free survival in accordance with the earlier studies. This may indicate that the aggressive chemotherapy employed for treatment of the predominantly ER-negative basal-like tumors increased survival duration in these patients relative to patients with tumors in the other subgroups. Thus, outcome for patients with basal-like tumors may not be as bad as indicated by earlier prognostic studies of patient populations that did not receive aggressive chemotherapy for progressive disease. Alternately, the differences may be due to differences in cohort selection. In either case, this result emphasizes the need to interpret the performance of molecular markers for patient stratification in the context of specific treatment regimens and in molecularly defined cohorts.

Second, we found that aggressively treated patients with high-level amplification had worse outcome than did patients without amplification (see Figure 3C). This is consistent with earlier CGH and single-locus analyses of associations of amplification with poor prognosis (Al-Kuraya et al., 2004; Blegen et al., 2003; Callagy et al., 2005; Gelsi-Boyer et al., 2005; Weber-Mangal et al., 2003). Moreover, the presence of high-level amplification was an indicator of poor outcome, even within patient subsets defined by expression profiling. This was particularly apparent

for luminal A tumors, as illustrated in Figure 3D, where patients whose tumors had high-level amplification at 8p11-12, 11q13-14, or 20q13 did significantly worse than patients without amplification. This suggests that stratification according to both expression level and copy number will identify patients that respond poorly to current therapeutic treatment strategies.

Mechanisms of disease progression

Our combined analyses of genome copy number and gene expression showed substantial differences in recurrent genome abnormality composition between tumors classified according to expression pattern and revealed that over 10% of the genes interrogated in this study had expression levels that were highly significantly associated with genome copy number changes. Most of the gene expression changes were associated with low-level changes in genome copy number, but 66 were deregulated by the high-level amplifications associated with poor outcome. These analyses provide insights into the etiology of breast cancer subtypes, suggest mechanisms by which the low-level copy number changes contribute to cancer pathogenesis, and identify a suite of genes that contribute to cancer pathophysiology.

Breast cancer subtypes

Figures 1 and 2 show that recurrent genome copy number aberrations differ substantially between tumors classified according to expression pattern as described previously (Perou et al., 1999). This is consistent with a model of cancer progression in which the expression subtype and genotype are determined by the cell type and stage of differentiation that survives telomere crisis and acquires sufficient proliferative advantage to achieve clonal dominance in the tumor (Chin et al., 2004). This model suggests that the genome CNA spectrum is selected to be most advantageous to the progression of the specific cell type that achieves immortality and clonal dominance. In this model, the recurrent genome CNA composition can be considered an independent subtype descriptor—much as genome CNA composition can be considered to be a cancer type descriptor (Knuutila et al., 2000). The independence of the genome CNA composition and basal and luminal expression subtypes is clear from Figure 4, which shows that the breast tumors divide into basal and luminal subtypes using unsupervised hierarchical clustering even after all transcripts showing associations with copy number are removed from the data set. Of course, the ERBB2 subtype is lost, since that subtype is strongly driven by ERBB2 amplification.

Low-level abnormalities

The most frequent low-level copy number changes were not associated with reduced survival duration, although some were associated with other markers usually associated with survival such as tumor size, nodal status, and grade (see Table 2). This raises the question of why the recurrent low-level CNAs are selected. GOstat analyses of the genes deregulated by these abnormalities showed that numerous genes involved in RNA and cellular metabolism were significantly upregulated by these events. Interestingly, we found these same GO classes to be significantly altered in a collection of breast cancer cell lines and in a study of ovarian cancer (W.-L.K., unpublished data). We also observed that many of the recurrent low-level aberrations matched the low-level copy number changes in the ZNF217-transfected human mammary epithelial cells that emerged after passage through telomere crisis having achieved clonal dominance in the culture (Chin et al., 2004; see Figure S2)—presumably because the aberrations they carried conferred a proliferative advantage. This suggests to us that the low-level CNAs are selected during early cancer formation because they increase basal metabolism, thereby providing a net survival/proliferative advantage to the cells that carry them. This idea is supported by a report that some of these same classes of genes were associated with proliferative fitness yeast (Deutschbauer et al., 2005). That study described analyses of proliferative fitness in the complete set of *Saccharomyces cerevisiae* heterozygous deletion strains and reported reduced growth rates for strains carrying deletions in genes involved in RNA metabolism and ribosome biogenesis and assembly.

High-level amplification

We found that high-level amplifications were associated with reduced survival duration and/or distant recurrence overall and within the luminal A expression subgroup. We identified 66 genes in these regions whose expression levels were correlated with copy number. GO analyses of those genes showed that they are involved in aspects of nucleic acid metabolism, protein modification, signaling, and the cell cycle and/or protein transport, and evidence is mounting that many if not most of these genes are functionally important in the cancers in which they are amplified and overexpressed (see Table 3). Indeed, published functional studies in model systems already have implicated eleven of these genes in diverse aspects of cancer pathophysiology. Six of these are encoded in the region of amplification at 8p11. These encode the RNA-binding protein LSM1 (Fraser et al., 2005), the receptor tyrosine kinase FGFR1 (Braun and Shannon, 2004), the cell-cycle-regulatory protein TACC1 (Still et al., 1999), the metalloproteinase ADAM9 (Mazzocca et al., 2005), the serine/threonine kinase IKBKB (Greten and Karin, 2004; Lam et al., 2005), and the DNA polymerase POLB (Clairmont et al., 1999). Functionally validated genes in the region of amplification at 11q13 include the cell-cycle-regulatory protein CCND1 (Hinds et al., 1994) and the growth factor FGF3 (Okunieff et al., 2003). Functionally important genes in the region of amplification at 17q include the transcription regulation protein PPARBP (Zhu et al., 2000), the receptor tyrosine kinase ERBB2 (Slamon et al., 1989), and the adaptor protein GRB7 (Tanaka et al., 2000), while the AKT-pathway-associated transcription factor ZNF217 (Huang et al., 2005; Nonet et al., 2001) and the RNA-binding protein REA1 (Babu et al., 2003) are functionally validated genes encoded in the region of amplification at 20q13. Further support for the functional importance of seven of these genes (*TACC1*, *ADAM9*, *IKBKB*, *POLB*, *CCND1*, *GRB7*, and *ZNF217*) in oncogenesis comes from the observation that they are within 100 Kbp of sites of recurrent tumorigenic viral integration in the mouse (Akagi et al., 2004), and three (*IKBKB*, *CCND1*, and *GRB7*) are within 10 Kbp of such a site. Taking proximity to a site of recurrent tumorigenic viral integration as evidence for a role in cancer genesis implicates an additional 13 genes or transcripts (see Table 3).

The biological roles of the genes deregulated by recurrent high-level amplification are diverse and vary between regions of amplification. For example, genes deregulated by amplification at 11q13 and 17q11-12 predominantly involved signaling and cell cycle regulation, while genes deregulated by amplification at 8p11-12 and 20q13 were of mixed function but were associated most frequently with aspects of nucleic acid metabolism. The predominance of genes involved in nucleic acid

Table 3. Functional characteristics of genes in recurrent amplicons associated with reduced survival duration in breast cancer

Gene	Ch	Mbp	p value, amplification	p value, disease free survival	p value, distant recurrence	Transcript description	Cancer function reference	Kbp to site of viral integration	Druggable?
SPFH2**	8	37.6	7.08E-07	0.053	0.003	chromosome 8 open reading frame 2			
PROSC**	8	37.7	2.28E-05	0.390	0.043	racemase and epimerase activity, energy metabolism			yes
BRF2**	8	37.8	1.20E-05	0.004	0.003	transcription factor regulating nucleic acid metabolism			
RAB11FIP1	8	37.8	7.77E-04	0.620	0.250	GTPase-activating protein involved in signal transduction			
ASH2L**	8	38.0	5.88E-06	0.036	0.002	DNA-binding protein involved in nucleic acid metabolism			
LSM1	8	38.0	6.79E-06	0.300	0.130	RNA-binding protein involved in nucleic acid metabolism	Fraser et al., 2005; Takahashi et al., 2002		
BAG4	8	38.1	8.73E-07	0.330	0.063	BCL2-associated chaperone protein involved in apoptosis	Gehrmann et al., 2005		
DDHD2**	8	38.1	4.40E-06	0.008	0.006	phospholipase involved in energy metabolism			
WHSC1L1	8	38.2	9.04E-06	0.760	0.730	nucleic acid binding			
FGFR1**	8	38.3	1.04E-04	0.025	0.540	receptor tyrosine kinase involved in signal transduction	Braun and Shannon, 2004; Ray et al., 2004		yes/ PD173074
TACC1**	8	38.7	6.72E-03	0.020	0.043	cell cycle control protein associated with signal transduction	Still et al., 1999	44.1/Plekha2	
ADAM9	8	38.9	1.91E-04	0.930	0.960	metalloproteinase associated with protein metabolism	Mazzocca et al., 2005	75/Plekha2	yes
GOLGA7	8	41.4	7.10E-05	0.140	0.170	integral membrane protein associated with transport			
SLD5	8	41.4	1.41E-03	0.780	0.460	unknown			
MYST3**	8	41.8	5.74E-05	0.006	0.022	transcription-regulatory protein involved in nucleic acid metabolism			
AP3M2**	8	42.0	4.43E-05	0.038	0.220	adapter protein associated with transport			
IKBKB**	8	42.1	7.73E-05	0.002	0.002	serine/threonine kinase associated with signal transduction	Greten and Karin, 2004; Lam et al., 2005	3.1/AK018683	yes/ PS-1145
POLB**	8	42.2	2.15E-04	0.001	0.008	DNA polymerase involved in nucleic acid metabolism	Clairmont et al., 1999	70.1/AK018683	
VDAC3**	8	42.3	9.93E-05	0.056	0.290	voltage-dependent anion channel associated with transport			
SLC20A2	8	42.3	1.98E-03	0.170	0.240	membrane transport protein			
THAP1**	8	42.7	7.13E-03	0.190	0.097	unknown			
FNTA**	8	42.9	3.13E-03	0.067	0.370	prenyltransferase associated with protein metabolism			yes
LOC441347	8	43.0	7.77E-04	0.180	0.810	unknown			
CCND1	11	69.2	1.50E-06	0.560	0.770	cell cycle control protein involved in signal transduction	Hinds et al., 1994	0.4/Fgf3	
FGF3	11	69.4	1.84E-03	0.920	0.420	growth factor involved in signal transduction	Okunieff et al., 2003		
FADD	11	70.0	7.42E-03	0.200	0.250	adapter molecule associated with signal transduction			
PPFIA1	11	70.0	1.53E-05	0.670	0.550	anchor protein associated with cell growth and/or maintenance			
CTTN*	11	70.0	2.69E-04	0.450	0.100	cytoskeletal protein associated with cell growth and/or maintenance			
NADSYN1	11	70.9	3.42E-04	0.290	0.990	unknown			
KRTAP5-9*	11	71.0	3.72E-03	0.035	0.050	cytoskeletal protein associated with cell growth and/or maintenance			
FOLR3	11	71.6	1.54E-03	0.730	0.490	cell surface receptor associated with signal transduction			
NEU3	11	74.4	9.73E-03	0.460	0.370	neuraminidase associated with protein metabolism			
N-PAC**	11	75.8	4.39E-03	0.110	0.038	protein kinase			
LHX1*	17	35.5	1.41E-03	0.250	0.018	transcription factor associated with nucleic acid metabolism			

Table 3. *Continued*

Gene	Ch	Mbp	p value, amplification	p value, disease free survival	p value, distant recurrence	Transcript description	Cancer function reference	Kbp to site of viral integration	Druggable?
ACACA	17	35.6	8.24E-03	0.850	0.850	carboxylase associated with energy metabolism			yes
DDX52	17	36.2	3.47E-04	0.300	0.560	RNA-binding protein associated with nucleic acid metabolism			
TBC1D3	17	36.7	5.25E-05	0.170	0.170	unknown			
SOCS7	17	36.9	4.00E-03	0.450	0.600	adapter molecule associated with signal transduction			
PCGF2	17	37.3	3.10E-04	0.760	0.850	transcription-regulatory protein associated with nucleic acid metabolism		5.4/Lasp1	
PSMB3	17	37.3	8.01E-03	0.390	0.810	ubiquitin proteasome system protein associated with protein metabolism		24.4/Lasp1	
PIP5K2B	17	37.3	5.07E-03	0.400	0.380	lipid kinase associated with signal transduction		47.5/Lasp1	
FLJ20291	17	37.3	3.14E-03	0.850	0.920	unknown		72.4/Lasp1	
PPARBP*	17	37.9	2.13E-04	0.089	0.260	transcription-regulatory protein associated with signal transduction	Zhu et al., 2000		
STARD3	17	38.2	3.40E-09	0.420	0.820	mitochondrial carrier protein associated with transport		52.1/Znfn1a3	
TCAP	17	38.2	1.26E-05	0.640	0.700	structural protein associated with cell growth and/or maintenance		23.1/Znfn1a3	
PNMT*	17	38.2	2.02E-06	0.630	0.010	methyltransferase associated with metabolism and energy		21.1/Znfn1a3	yes
PERLD1	17	38.2	3.41E-09	0.930	0.840	membrane protein of unknown function		18.2/Znfn1a3	
ERBB2	17	38.2	3.41E-09	0.110	0.560	receptor tyrosine kinase associated with signal transduction	Slamon et al., 1989		yes/ trastuzumab, lapatinib
GRB7*	17	38.3	7.28E-08	0.044	0.300	adapter molecule associated with signal transduction	Tanaka et al., 2000	10.8/Znfn1a3	
GSDML	17	38.4	8.36E-06	0.710	0.690	unknown		48.8/Znfn1a3	
PSMD3	17	38.5	4.25E-03	0.250	0.510	ubiquitin proteasome system protein associated with protein metabolism		32.8/Znfn1a3	
NR1D1	17	38.6	1.28E-03	0.210	0.750	nuclear receptor associated with signal transduction		73.4/Cdc6	yes
ZNF217	20	52.9	5.02E-06	0.650	0.650	transcription factor associated with signal transduction	Nonet et al., 2001	39.3/Zfp217	
BCAS1	20	53.2	4.93E-03	0.290	0.140	unknown		70.9/Zpf217	
CSTF1	20	55.7	7.15E-03	0.150	0.330	pre-mRNA processing			
RAE1	20	56.6	3.56E-05	0.360	0.420	RNA-binding protein associated with nucleic acid metabolism	Babu et al., 2003		
RNPC1	20	56.6	1.19E-03	0.750	0.830	RNA-binding protein associated with nucleic acid metabolism			
PCK1	20	56.8	9.78E-03	0.250	0.330	phosphotransferase associated with energy and metabolism			
TMEPAI*	20	56.9	1.21E-04	0.085	0.077	unknown			
RAB22A	20	57.6	3.15E-05	0.990	0.340	GTPase associated with signal transduction			
VAPB	20	57.6	3.78E-05	0.360	0.260	membrane transport protein			
STX16	20	57.9	2.63E-05	0.220	0.790	transport/cargo protein			
NPEPL1	20	57.9	3.35E-05	0.270	0.800	aminopeptidase associated with protein metabolism			
GNAS**	20	58.1	6.60E-03	0.052	0.058	G protein associated with signal transduction			
TH1L	20	58.2	1.14E-04	0.530	0.800	transcription-regulatory protein associated with nucleic acid metabolism		36.7/Thil	
C20orf45	20	58.3	6.29E-04	0.970	0.790	unknown		88.7/Th1l	

Functional annotation was based on the Human Protein Reference Database (http://hprd.org/). Genes marked with an asterisk are associated with reduced survival duration or distant recurrence when overexpressed in nonamplifying tumors. Genes marked with two asterisks are significantly associated with reduced survival duration or distant recurrence ($p < 0.05$) when downregulated in nonamplifying tumors. Distances to sites of recurrent viral integration were determined from published information (Akagi et al., 2004). The last column identifies genes that have predicted protein folding characteristics that suggest that they might be druggable (Russ and Lampel, 2005).

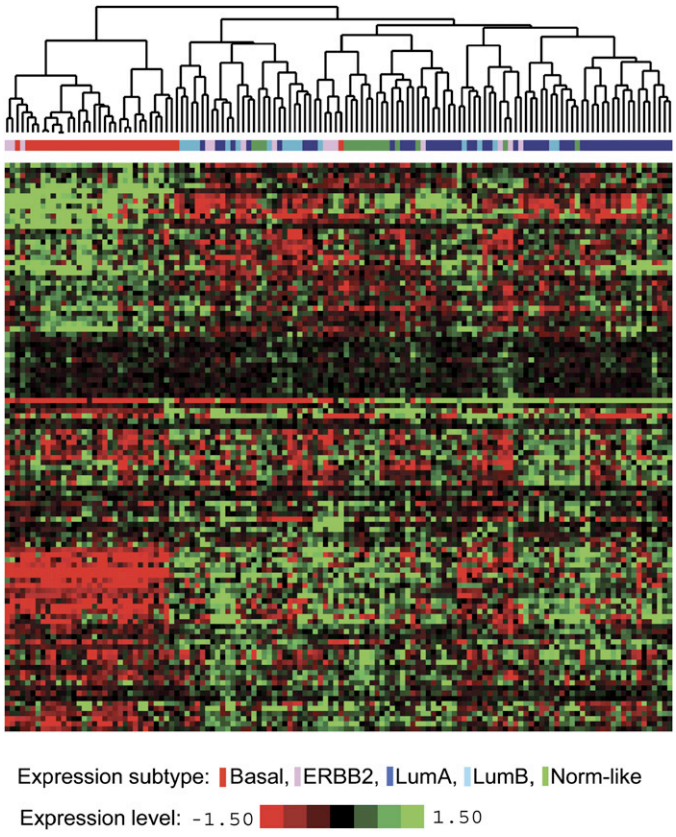

Figure 4. Results of unsupervised hierarchical clustering of 130 breast tumors using intrinsically variable gene expression but excluding any transcripts whose levels were significantly associated with genome copy number

Red indicates increased expression, and green indicates reduced expression. An annotated version is provided as Figure S3.

metabolism in the region of amplification at 8p11-12 was especially strong. Interestingly, the region of recurrent amplification at 8p11-12 described above was *reduced* in copy number in some tumors, and this event also was associated with poor outcome. This raises the possibility that poor clinical outcome in tumors with 8p11-12 abnormalities is due to increased genome instability/mutagenesis resulting from either up- or downregulation of genes encoded in this region. This concept is supported by studies in yeast showing that up- or downregulation of genes involved in chromosome integrity and segregation can produce similar instability phenotypes (Ouspenski et al., 1999).

Therapeutic targets

The 66 genes we found to be deregulated by the high-level amplifications associated with poor outcome are particularly interesting as therapeutic targets for treatment of patients that are refractory to current therapies. Small-molecule or antibody-based inhibitors have already been developed for *FGFR1* (PD173074; Ray et al., 2004), *IKBKB* (PS-1145; Lam et al., 2005), and *ERBB2* (Trastuzumab; Vogel et al., 2002), and six others (*PROCC*, *ADAM9*, *FNTA*, *ACACA*, *PNMT*, and *NR1D1*) are considered to be druggable based on the presence of predicted protein folds that favor interactions with drug-like compounds (Russ and Lampel, 2005). Taking *ERBB2* as the paradigm (recurrently amplified, overexpressed, associated with outcome and with demonstrated functional importance in cancer) suggests *FGFR1*, *TACC1*, *ADAM9*, *IKBKB*, *PNMT*, and *GRB7* as high-priority therapeutic targets in these regions of amplification.

Experimental procedures

Tumor characteristics

Frozen tissue from UC San Francisco and the California Pacific Medical Center collected between 1989 and 1997 was used for this study. Tissues were collected under IRB-approved protocols with patient consent. Tissues were collected, frozen over dry ice within 20 min of resection, and stored at −80°C. An H&E section of each tumor sample was reviewed, and the frozen block was manually trimmed to remove normal and necrotic tissue from the periphery. Clinical follow-up was available with a median time of 6.6 years overall and 8 years for censored patients. Tumors were predominantly early stage (83% stage I and II) with an average diameter of 2.6 cm. About half of the tumors were node positive, 67% were estrogen receptor positive, 60% received tamoxifen, and half received adjuvant chemotherapy (typically adriamycin and cytoxan). Clinical characteristics of the individual tumors are provided together with expression and array CGH profiles in the CaBIG repository and at http://cancer.lbl.gov/breastcancer/data.php.

Array CGH

Each sample was analyzed using Scanning and OncoBAC arrays. Scanning arrays were comprised of 2464 BACs selected at approximately megabase intervals along the genome as described previously (Hodgson et al., 2001; Snijders et al., 2001). OncoBAC arrays were comprised of 960 P1, PAC, or BAC clones. About three-quarters of the clones on the OncoBAC arrays contained genes and STSs implicated in cancer development or progression. All clones were printed in quadruplicate. DNA samples for array CGH were labeled generally as described previously (Hackett et al., 2003; Hodgson et al., 2001; Snijders et al., 2001). Briefly, 500 ng each of cancer and normal female genomic DNA sample was labeled by random priming with CY3- and CY5-dUTP, respectively; denatured; and hybridized with unlabeled Cot-1 DNA to CGH arrays. After hybridization, the slides were washed and imaged using a 16-bit CCD camera through CY3, CY5, and DAPI filters (Pinkel et al., 1998).

Expression profiling

Expression profiling was accomplished using the Affymetrix High Throughput Array (HTA) GeneChip system, in which target preparations, washing, and staining were carried out in a 96-well format. Detailed methods are described in the Supplemental Data.

Statistical considerations

Data processing

Array CGH data image analyses were performed as described previously (Jain et al., 2002). In this process, an array probe was assigned a missing value for an array if there were fewer than two valid replicates or the standard deviation of the replicates exceeded 0.2. Array probes missing in more than 50% of samples in OncoBAC or scanning array data sets were excluded in the corresponding set. Array probes representing the same DNA sequence were averaged within each data set and then between the two data sets. Finally, the two data sets were combined, and the array probes missing in more than 25% of the samples, unmapped array probes, and probes mapped to chromosome Y were eliminated. The final data set contained 2149 unique probes. For Affymetrix data, multichip robust normalization was performed using RMA software (Irizarry et al., 2003). Transcripts assessed on the arrays were classified into two groups using Gaussian model-based clustering by considering the joint distribution of the median and standard deviation of each probe set across samples. During this process, computational demands were reduced by randomly sampling and clustering 2000 probe intensities using *mclust* (Yeung et al., 2001, 2004) with two clusters and unequal variance. Next, the remaining probe intensities were classified into the newly created clusters using linear discriminant analysis. The cluster containing probe intensities with smaller mean and variance was defined as "not expressed," and the second cluster was defined as "expressed."

Characterizing copy number changes

The sample profiles were segmented into the levels of equal copy number common to the whole genome, and the copy number transitions,

amplifications, and frequency of alterations were determined using previously described methodologies (Snijders et al., 2003; Fridlyand et al., 2006). The detailed approaches are described in the Supplemental Data.

Clustering of genome copy number profiles

Genome copy number profiles were clustered using smoothed imputed data with outliers present. Agglomerative hierarchical clustering with Pearson's correlation as a similarity measure and the Ward method to minimize sum of variances were used to produce compact spherical clusters (Hartigan, 1975). The number of groups was assessed qualitatively by considering the shape of the clustering dendogram.

Expression subtype assignment

Tumors were classified according to expression phenotype (basal, ERBB2, luminal A, luminal B, and normal-like) by assigning each tumor to the subtype of the cluster defined by hierarchical clustering of expression profiles for 122 samples published by Sorlie et al. (2003) to which it had the highest Pearson's correlation. The correlation was computed using the subset of Stanford intrinsically variable genes common to both data sets. For details, refer to the Supplemental Data.

Association of copy number with survival

Stage 4 samples were excluded from all the outcome-related analyses, and disease-specific survival and time to distant recurrence were used as the two endpoints. Significance of the standardized regression coefficient Cox-proportional model was used to determine clinical (univariate and multivariate analyses) and genomic variables (individual clones, instability summary measures, and recurrent amplicon status) associated with outcome. p values for individual clones were adjusted using FDR. The significance was declared at $p < 0.05$. For details, see the Supplemental Data.

Association of copy number with expression

The presence of an overall dosage effect was assessed by subdividing each chromosomal arm into nonoverlapping 20 Mb bins and computing the average of cross-Pearson's-correlations for all gene transcript-BAC probe pairs that mapped to that bin. We also calculated Pearson's correlations and corresponding p values between expression level and copy number for each gene transcript. Each transcript was assigned an observed copy number of the nearest mapped BAC array probe. Eighty percent of gene transcripts had a nearest clone within 1 Mbp, and 50% had a clone within 400 Kbp. Correlation between expression and copy number was only computed for the gene transcripts whose absolute assigned copy number exceeded 0.2 in at least five samples. This was done to avoid spurious correlations in the absence of real copy number changes. We used conservative Holm p value adjustment to correct for multiple testing. Gene transcripts with an adjusted p value <0.05 were considered to have expression levels that were highly significantly affected by gene dosage. This corresponded to a minimum Pearson's correlation of 0.44.

Associations of transcription and CNA in regions of amplification with outcome in tumors without particular amplicons

We assessed the associations of levels of transcripts in regions of amplifications with survival or distant recurrence in tumors without amplifications in order to find genes that might contribute to progression when deregulated by mechanisms other than amplification (e.g., we assessed associations between expression levels of the genes mapping to the 8p11-12 amplicon and survival in samples without 8p11-12 amplification). We performed separate Cox-proportional regressions for disease-specific survival and distant recurrence. Stage 4 samples were excluded from all analyses.

Testing for functional enrichment

We used the gene ontology statistics tool GoStat (Beissbarth and Speed, 2004) to test whether the gene transcripts with the strongest dosage effects were enriched for particular functional groups. The p values were adjusted using false discovery rate. The categories were considered significantly overrepresented if the FDR-adjusted p value was less than 0.001. Since expressed genes were significantly more likely to show dosage effects than nonexpressed genes (p value < 2.2E-16, Wilcoxon rank-sum test), GoStat comparisons were performed only for expressed genes. Specifically, GO categories for 1734 expressed probes with significant dosage effect (Holm p value < 0.05) were compared with those for 3026 expressed probes with no dosage effect (Pearson's correlation < 0.1).

Microarray data

The raw data for expression profiling are available at ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) with accession number E-TABM-158.

Clinical characteristics of the individual tumors as well as array CGH and expression profiles are available in the CaBIG repository (http://caarraydb.nci.nih.gov/caarray/publicExperimentDetailAction.do?expId=1015897589973255), at http://cancer.lbl.gov/breastcancer/data.php, and in the Supplemental Data.

Supplemental data

The Supplemental Data include Supplemental Experimental Procedures, three supplemental figures, and three supplemental tables and can be found with this article online at http://www.cancercell.org/cgi/content/full/10/6/529/DC1/.

Acknowledgments

This work was supported by the NIH (CA58207, CA90421, and CA101359), the Office of Health and Environmental Research of the U.S. Department of Energy (contract DE-AC03-76SF00098), and the Avon Foundation. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. For full disclaimer, see http://www-library.lbl.gov/public/tmRco/howto/RcoBerkeleyLabDisclaimer.htm.

Received: May 3, 2006
Revised: August 19, 2006
Accepted: October 6, 2006
Published: December 11, 2006

References

Akagi, K., Suzuki, T., Stephens, R.M., Jenkins, N.A., and Copeland, N.G. (2004). RTCGD: Retroviral tagged cancer gene database. Nucleic Acids Res. *32*, D523–D527.

Al-Kuraya, K., Schraml, P., Torhorst, J., Tapia, C., Zaharieva, B., Novotny, H., Spichtin, H., Maurer, R., Mirlacher, M., Kochli, O., et al. (2004). Prognostic relevance of gene amplifications and coamplifications in breast cancer. Cancer Res. *64*, 8534–8540.

Albertson, D.G., Collins, C., McCormick, F., and Gray, J.W. (2003). Chromosome aberrations in solid tumors. Nat. Genet. *34*, 369–376.

Babu, J.R., Jeganathan, K.B., Baker, D.J., Wu, X., Kang-Decker, N., and van Deursen, J.M. (2003). Rae1 is an essential mitotic checkpoint regulator that cooperates with Bub3 to prevent chromosome missegregation. J. Cell Biol. *160*, 341–353.

Barlund, M., Tirkkonen, M., Forozan, F., Tanner, M.M., Kallioniemi, O., and Kallioniemi, A. (1997). Increased copy number at 17q22-q24 by CGH in breast cancer is due to high-level amplification of two separate regions. Genes Chromosomes Cancer *20*, 372–376.

Barlund, M., Monni, O., Kononen, J., Cornelison, R., Torhorst, J., Sauter, G., Kallioniemi, O.-P., and Kallioniemi, A. (2000). Multiple genes at 17q23 undergo amplification and overexpression in breast cancer. Cancer Res. *60*, 5340–5344.

Baylin, S.B., and Herman, J.G. (2000). DNA hypermethylation in tumorigenesis: Epigenetics joins genetics. Trends Genet. *16*, 168–174.

Bergamaschi, A., Kim, Y.H., Wang, P., Sorlie, T., Hernandez-Boussard, T., Lonning, P.E., Tibshirani, R., Borresen-Dale, A.L., and Pollack, J.R. (2006). Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. Genes Chromosomes Cancer *45*, 1033–1040.

Beissbarth, T., and Speed, T.P. (2004). GOstat: Find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics *20*, 1464–1465.

Blegen, H., Will, J.S., Ghadimi, B.M., Nash, H.P., Zetterberg, A., Auer, G., and Ried, T. (2003). DNA amplifications and aneuploidy, high proliferative activity and impaired cell cycle control characterize breast carcinomas with poor prognosis. Anal. Cell. Pathol. *25*, 103–114.

Braun, B.S., and Shannon, K. (2004). The sum is greater than the FGFR1 partner. Cancer Cell *5*, 203–204.

Callagy, G., Pharoah, P., Chin, S.F., Sangan, T., Daigo, Y., Jackson, L., and Caldas, C. (2005). Identification and validation of prognostic markers in breast cancer with the complementary use of array-CGH and tissue microarrays. J. Pathol. *205*, 388–396.

Cheng, K.W., Lahad, J.P., Kuo, W.L., Lapuk, A., Yamada, K., Auersperg, N., Liu, J., Smith-McCune, K., Lu, K.H., Fishman, D., et al. (2004). The RAB25 small GTPase determines aggressiveness of ovarian and breast cancers. Nat. Med. *10*, 1251–1256.

Chin, K., de Solorzano, C.O., Knowles, D., Jones, A., Chou, W., Rodriguez, E.G., Kuo, W.L., Ljung, B.M., Chew, K., Myambo, K., et al. (2004). In situ analyses of genome instability in breast cancer. Nat. Genet. *36*, 984–988.

Clairmont, C.A., Narayanan, L., Sun, K.W., Glazer, P.M., and Sweasy, J.B. (1999). The Tyr-265-to-Cys mutator mutant of DNA polymerase beta induces a mutator phenotype in mouse LN12 cells. Proc. Natl. Acad. Sci. USA *96*, 9580–9585.

Deutschbauer, A.M., Jaramillo, D.F., Proctor, M., Kumm, J., Hillenmeyer, M.E., Davis, R.W., Nislow, C., and Giaever, G. (2005). Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. Genetics *169*, 1915–1925.

Esteva, F.J., Sahin, A.A., Cristofanilli, M., Coombes, K., Lee, S.J., Baker, J., Cronin, M., Walker, M., Watson, D., Shak, S., and Hortobagyi, G.N. (2005). Prognostic role of a multigene reverse transcriptase-PCR assay in patients with node-negative breast cancer not receiving adjuvant systemic therapy. Clin. Cancer Res. *11*, 3315–3319.

Fraser, M.M., Watson, P.M., Fraig, M.M., Kelley, J.R., Nelson, P.S., Boylan, A.M., Cole, D.J., and Watson, D.K. (2005). CaSm-mediated cellular transformation is associated with altered gene expression and messenger RNA stability. Cancer Res. *65*, 6228–6236.

Fridlyand, J., Snijders, A.M., Ylstra, B., Li, H., Olshen, A., Segraves, R., Dairkee, S., Tokuyasu, T., Ljung, B.M., Jain, A.N., et al. (2006). Breast tumor copy number aberration phenotypes and genomic instability. BMC Cancer *6*, 96.

Gehrmann, M., Marienhagen, J., Eichholtz-Wirth, H., Fritz, E., Ellwart, J., Jaattela, M., Zilch, T., and Multhoff, G. (2005). Dual function of membrane-bound heat shock protein 70 (Hsp70), Bag-4, and Hsp40: Protection against radiation-induced effects and target structure for natural killer cells. Cell Death Differ. *12*, 38–51.

Gelsi-Boyer, V., Orsetti, B., Cervera, N., Finetti, P., Sircoulomb, F., Rouge, C., Lasorsa, L., Letessier, A., Ginestier, C., Monville, F., et al. (2005). Comprehensive profiling of 8p11-12 amplification in breast cancer. Mol. Cancer Res. *3*, 655–667.

Gianni, L., Zambetti, M., Clark, K., Baker, J., Cronin, M., Wu, J., Mariani, G., Rodriguez, J., Carcangiu, M., Watson, D., et al. (2005). Gene expression profiles in paraffin-embedded core biopsy tissue predict response to chemotherapy in women with locally advanced breast cancer. J. Clin. Oncol. *23*, 7265–7277.

Greten, F.R., and Karin, M. (2004). The IKK/NF-κB activation pathway—A target for prevention and treatment of cancer. Cancer Lett. *206*, 193–199.

Hackett, C.S., Hodgson, J.G., Law, M.E., Fridlyand, J., Osoegawa, K., de Jong, P.J., Nowak, N.J., Pinkel, D., Albertson, D.G., Jain, A., et al. (2003). Genome-wide array CGH analysis of murine neuroblastoma reveals distinct genomic aberrations which parallel those in human tumors. Cancer Res. *63*, 5266–5273.

Hanahan, D., and Weinberg, R.A. (2000). The hallmarks of cancer. Cell *100*, 57–70.

Hartigan, J.A. (1975). Clustering Algorithms (New York: Wiley).

Hinds, P.W., Dowdy, S.F., Eaton, E.N., Arnold, A., and Weinberg, R.A. (1994). Function of a human cyclin gene as an oncogene. Proc. Natl. Acad. Sci. USA *91*, 709–713.

Hodgson, G., Hager, J.H., Volik, S., Hariono, S., Wernick, M., Moore, D., Nowak, N., Albertson, D.G., Pinkel, D., Collins, C., et al. (2001). Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. Nat. Genet. *29*, 459–464.

Huang, G., Krig, S., Kowbel, D., Xu, H., Hyun, B., Volik, S., Feuerstein, B., Mills, G.B., Stokoe, D., Yaswen, P., and Collins, C. (2005). ZNF217 suppresses cell death associated with chemotherapy and telomere dysfunction. Hum. Mol. Genet. *14*, 3219–3225.

Hyman, E., Kauraniemi, P., Hautaniemi, S., Wolf, M., Mousses, S., Rozenblum, E., Ringner, M., Sauter, G., Monni, O., Elkahloun, A., et al. (2002). Impact of DNA amplification on gene expression patterns in breast cancer. Cancer Res. *62*, 6240–6245.

Irizarry, R., Bolstad, B., Collin, F., Cope, L., Hobbs, B., and Speed, T. (2003). Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. *31*, e15.

Isola, J.J., Kallioniemi, O.P., Chu, L.W., Fuqua, S.A., Hilsenbeck, S.G., Osborne, C.K., and Waldman, F.M. (1995). Genetic aberrations detected by comparative genomic hybridization predict outcome in node-negative breast cancer. Am. J. Pathol. *147*, 905–911.

Jain, A.N., Chin, K., Borresen-Dale, A.L., Erikstein, B.K., Eynstein Lonning, P., Kaaresen, R., and Gray, J.W. (2001). Quantitative analysis of chromosomal CGH in human breast tumors associates copy number abnormalities with p53 status and patient survival. Proc. Natl. Acad. Sci. USA *98*, 7952–7957.

Jain, A.N., Tokuyasu, T.A., Snijders, A.M., Segraves, R., Albertson, D.G., and Pinkel, D. (2002). Fully automatic quantification of microarray image data. Genome Res. *12*, 325–332.

Jones, P.A. (2005). Overview of cancer epigenetics. Semin. Hematol. *42*, S3–S8.

Kallioniemi, O.P., Kallioniemi, A., Kurisu, W., Thor, A., Chen, L.C., Smith, H.S., Waldman, F.M., Pinkel, D., and Gray, J.W. (1992). ERBB2 amplification in breast cancer analyzed by fluorescence in situ hybridization. Proc. Natl. Acad. Sci. USA *89*, 5321–5325.

Kallioniemi, A., Kallioniemi, O.P., Piper, J., Tanner, M., Stokke, T., Chen, L., Smith, H.S., Pinkel, D., Gray, J.W., and Waldman, F.M. (1994). Detection and mapping of amplified DNA sequences in breast cancer by comparative genomic hybridization. Proc. Natl. Acad. Sci. USA *91*, 2156–2160.

Kauraniemi, P., Barlund, M., Monni, O., and Kallioniemi, A. (2001). New amplified and highly expressed genes discovered in the ERBB2 amplicon in breast cancer by cDNA microarrays. Cancer Res. *61*, 8235–8240.

Kauraniemi, P., Kuukasjarvi, T., Sauter, G., and Kallioniemi, A. (2003). Amplification of a 280-kilobase core region at the ERBB2 locus leads to activation of two hypothetical proteins in breast cancer. Am. J. Pathol. *163*, 1979–1984.

Knuutila, S., Autio, K., and Aalto, Y. (2000). Online access to CGH data of DNA sequence copy number changes. Am. J. Pathol. *157*, 689.

Lam, L.T., Davis, R.E., Pierce, J., Hepperle, M., Xu, Y., Hottelet, M., Nong, Y., Wen, D., Adams, J., Dang, L., and Staudt, L.M. (2005). Small molecule inhibitors of IκB kinase are selectively toxic for subgroups of diffuse large B-cell lymphoma defined by gene expression profiling. Clin. Cancer Res. *11*, 28–40.

Loo, L.W., Grove, D.I., Williams, E.M., Neal, C.L., Cousens, L.A., Schubert, E.L., Holcomb, I.N., Massa, H.F., Glogovac, J., Li, C.I., et al. (2004). Array comparative genomic hybridization analysis of genomic alterations in breast cancer subtypes. Cancer Res. *64*, 8541–8549.

Mazzocca, A., Coppari, R., De Franco, R., Cho, J.Y., Libermann, T.A., Pinzani, M., and Toker, A. (2005). A secreted form of ADAM9 promotes carcinoma invasion through tumor-stromal interactions. Cancer Res. *65*, 4728–4738.

Naylor, T.L., Greshock, J., Wang, Y., Colligon, T., Yu, Q.C., Clemmer, V., Zaks, T.Z., and Weber, B.L. (2005). High resolution genomic analysis of sporadic breast cancer using array-based comparative genomic hybridization. Breast Cancer Res. *7*, R1186–R1198.

Neve, R.M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F.L., Fevr, T., Clark, L., Bayani, N., Coppe, J.-P., Tong, F., et al. (2006). A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. Cancer Cell *10*, this issue, 515–527.

Nonet, G., Stampfer, M., Chin, K., Gray, J.W., Collins, C., and Yaswen, P. (2001). The ZNF217 gene amplified in breast cancers promotes immortalization of human mammary epithelial cells. Cancer Res. *61*, 1250–1254.

Okunieff, P., Fenton, B.M., Zhang, L., Kern, F.G., Wu, T., Greg, J.R., and Ding, I. (2003). Fibroblast growth factors (FGFS) increase breast tumor growth rate, metastases, blood flow, and oxygenation without significant change in vascular density. Adv. Exp. Med. Biol. *530*, 593–601.

Ouspenski, I.I., Elledge, S.J., and Brinkley, B.R. (1999). New yeast genes important for chromosome integrity and segregation identified by dosage effects on genome stability. Nucleic Acids Res. *27*, 3001–3008.

Perou, C.M., Jeffrey, S.S., van de Rijn, M., Rees, C.A., Eisen, M.B., Ross, D.T., Pergamenschikov, A., Williams, C.F., Zhu, S.X., Lee, J.C., et al. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc. Natl. Acad. Sci. USA *96*, 9212–9217.

Perou, C.M., Sorlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., et al. (2000). Molecular portraits of human breast tumours. Nature *406*, 747–752.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.L., Chen, C., Zhai, Y., et al. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat. Genet. *20*, 207–211.

Pollack, J.R., Perou, C.M., Alizadeh, A.A., Eisen, M.B., Pergamenschikov, A., Williams, C.F., Jeffrey, S.S., Botstein, D., and Brown, P.O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat. Genet. *23*, 41–46.

Pollack, J.R., Sorlie, T., Perou, C.M., Rees, C.A., Jeffrey, S.S., Lonning, P.E., Tibshirani, R., Botstein, D., Borresen-Dale, A.L., and Brown, P.O. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Proc. Natl. Acad. Sci. USA *99*, 12963–12968.

Press, M.F., Sauter, G., Bernstein, L., Villalobos, I.E., Mirlacher, M., Zhou, J.Y., Wardeh, R., Li, Y.T., Guzman, R., Ma, Y., et al. (2005). Diagnostic evaluation of HER-2 as a molecular target: An assessment of accuracy and reproducibility of laboratory testing in large, prospective, randomized clinical trials. Clin. Cancer Res. *11*, 6598–6607.

Ramaswamy, S., Ross, K.N., Lander, E.S., and Golub, T.R. (2003). A molecular signature of metastasis in primary solid tumors. Nat. Genet. *33*, 49–54.

Ray, M.E., Yang, Z.Q., Albertson, D., Kleer, C.G., Washburn, J.G., Macoska, J.A., and Ethier, S.P. (2004). Genomic and expression analysis of the 8p11-12 amplicon in human breast cancer cell lines. Cancer Res. *64*, 40–47.

Reyal, F., Stransky, N., Bernard-Pierrot, I., Vincent-Salomon, A., de Rycke, Y., Elvin, P., Cassidy, A., Graham, A., Spraggon, C., Desille, Y., et al. (2005). Visualizing chromosomes as transcriptome correlation maps: Evidence of chromosomal domains containing co-expressed genes—A study of 130 invasive ductal breast carcinomas. Cancer Res. *65*, 1376–1383.

Russ, A.P., and Lampel, S. (2005). The druggable genome: An update. Drug Discov. Today *10*, 1607–1610.

Slamon, D.J., Godolphin, W., Jones, L.A., Holt, J.A., Wong, S.G., Keith, D.E., Levin, W.J., Stuart, S.G., Udove, J., Ullrich, A., et al. (1989). Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. Science *244*, 707–712.

Snijders, A.M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A.K., Huey, B., Kimura, K., et al. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. Nat. Genet. *29*, 263–264.

Snijders, A.M., Fridlyand, J., Mans, D.A., Segraves, R., Jain, A.N., Pinkel, D., and Albertson, D.G. (2003). Shaping of tumor and drug-resistant genomes by instability and selection. Oncogene *22*, 4370–4379.

Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., Dohner, H., Cremer, T., and Lichter, P. (1997). Matrix-based comparative genomic hybridization: Biochips to screen for genomic imbalances. Genes Chromosomes Cancer *20*, 399–407.

Sorlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci. USA *98*, 10869–10874.

Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J.S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., et al. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc. Natl. Acad. Sci. USA *100*, 8418–8423.

Still, I.H., Hamilton, M., Vince, P., Wolfman, A., and Cowell, J.K. (1999). Cloning of TACC1, an embryonically expressed, potentially transforming coiled coil containing gene, from the 8p11 breast cancer amplicon. Oncogene *18*, 4032–4038.

Takahashi, S., Suzuki, S., Inaguma, S., Cho, Y.M., Ikeda, Y., Hayashi, N., Inoue, T., Sugimura, Y., Nishiyama, N., Fujita, T., et al. (2002). Down-regulation of Lsm1 is involved in human prostate cancer progression. Br. J. Cancer *86*, 940–946.

Tanaka, S., Sugimachi, K., Kawaguchi, H., Saeki, H., Ohno, S., and Wands, J.R. (2000). Grb7 signal transduction protein mediates metastatic progression of esophageal carcinoma. J. Cell. Physiol. *183*, 411–415.

Tanner, M.M., Tirkkonen, M., Kallioniemi, A., Collins, C., Stokke, T., Karhu, R., Kowbel, D., Shadravan, F., Hintz, M., Kuo, W.L., et al. (1994). Increased copy number at 20q13 in breast cancer: Defining the critical region and exclusion of candidate genes. Cancer Res. *54*, 4257–4260.

van de Vijver, M.J., He, Y.D., van 't Veer, L.J., Dai, H., Hart, A.A., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J., et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. N. Engl. J. Med. *347*, 1999–2009.

van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature *415*, 530–536.

Vogel, C.L., Cobleigh, M.A., Tripathy, D., Gutheil, J.C., Harris, L.N., Fehrenbacher, L., Slamon, D.J., Murphy, M., Novotny, W.F., Burchmore, M., et al. (2002). Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. J. Clin. Oncol. *20*, 719–726.

Weber-Mangal, S., Sinn, H.P., Popp, S., Klaes, R., Emig, R., Bentz, M., Mansmann, U., Bastert, G., Bartram, C.R., and Jauch, A. (2003). Breast cancer in young women ($\leq$35 years): Genomic aberrations detected by comparative genomic hybridization. Int. J. Cancer *107*, 583–592.

Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E., and Ruzzo, W.L. (2001). Model-based clustering and data transformations for gene expression data. Bioinformatics *17*, 977–987.

Yeung, K.Y., Medvedovic, M., and Bumgarner, R.E. (2004). From co-expression to co-regulation: How many microarray experiments do we need? Genome Biol. *5*, R48.

Yi, Y., Mirosevich, J., Shyr, Y., Matusik, R., and George, A.L., Jr. (2005). Coupled analysis of gene expression and chromosomal location. Genomics *85*, 401–412.

Zhu, Y., Kan, L., Qi, C., Kanwar, Y.S., Yeldandi, A.V., Rao, M.S., and Reddy, J.K. (2000). Isolation and characterization of peroxisome proliferator-activated receptor (PPAR) interacting protein (PRIP) as a coactivator for PPAR. J. Biol. Chem. *275*, 13510–13516.

Accession numbers

The raw data for expression profiling are available at ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) with accession number E-TABM-158.